

A P P L I C A T I O N

for

UNITED STATES LETTERS PATENT

on

METHODS FOR RAPID AND QUANTITATIVE PROTEOME ANALYSIS

by

Ruedi Aebersold

and

David R. Goodlett

CERTIFICATE OF MAILING BY "EXPRESS MAIL"

"EXPRESS MAIL" MAILING LABEL NUMBER: EL 690 156 345 US

DATE OF DEPOSIT: December 26, 2000

I HEREBY CERTIFY THAT THIS PAPER OR FEE IS BEING
DEPOSITED WITH THE UNITED STATES POSTAL SERVICE
"EXPRESS MAIL POST OFFICE TO ADDRESSEE" SERVICE UNDER
37 C.F.R. 1.10 ON THE DATE INDICATED ABOVE, AND IS
ADDRESSED TO: BOX PATENT APPLICATION,
COMMISSIONER FOR PATENTS, WASHINGTON, D.C. 20231.

RICHARD MONZON

(TYPED OR PRINTED NAME OR PERSON MAILING PAPER OR FEE)

(SIGNATURE OF PERSON MAILING PAPER OR FEE)

Sheets of Drawings: 2

Docket No.: P-IS 4534

Attorneys

CAMPBELL AND FLORES

4370 La Jolla Village Drive, Suite 700
San Diego, California 92122

USPTO Customer No. 23601

METHODS FOR RAPID AND QUANTITATIVE PROTEOME ANALYSIS

This invention was made with government support under grant number R33 CA84698-01 awarded by National Cancer Institute at the National Institutes of Health. The United States Government has certain rights in this invention.

BACKGROUND OF THE INVENTION

This invention relates generally to proteome analysis and, more specifically, to methods of identifying and/or quantifying a protein or proteins that is contained in a mixture of proteins.

The classical biochemical approach to study biological processes has been based on the purification to homogeneity by sequential fractionation and assay cycles of the specific activities that constitute a process, the detailed structural, functional and regulatory analysis of each isolated component, and the reconstitution of the process from the isolated components. The Human Genome Project and other genome sequencing programs are turning out in rapid succession the complete genome sequences of specific species and, thus, in principle the amino acid sequence of every protein potentially encoded by that species. It is to be expected that this information resource unprecedented in the history of biology will enhance traditional research methods and catalyze progress in fundamentally different research paradigms, one of which is Proteomics.

Efforts to sequence the entire human genome along with the genomes of a number of other species have been extraordinarily successful. The genomes of 46 microbial species (TIGR Microbial Database; www.tigr.org) have been
5 completed and the genomes of over one hundred twenty other microbial species are in the process of being sequenced. Additionally, the more complex genomes of eukaryotes, in particular those of the genetically well characterized unicellular organism *Saccharomyces cerevisiae* and the
10 multicellular species *Caenorhabditis elegans* and *Drosophila melanogaster* have been sequenced completely. Furthermore, "draft sequence" of the rice genome has been published, and completion of the human and *Arabidopsis* genomes are imminent. Even in the absence of complete genomic
15 sequences, rich DNA sequence databases have been made publicly available, including those containing over 2.1 million human and over 1.2 million murine expressed sequence tags (ESTs).

ESTs are stretches of approximately 300 to 500
20 contiguous nucleotides representing partial gene sequences that are being generated by systematic single pass sequencing of the clones in cDNA libraries. On the timescale of most biological processes, with the notable exception of evolution, the genomic DNA sequence can be
25 viewed as static, and a genomic sequence database therefore represents an information resource akin to a library. Intensive efforts are underway to assign "function" to individual sequences in sequence databases. This is attempted by the computational analysis of linear sequence
30 motifs or higher order structural motifs that indicate a statistically significant similarity of a sequence to a

family of sequences with known function, or by other means such as comparison of homologous protein functions across species. Other methods have also been used to determine function of individual sequences, including experimental methods such as gene knockouts and suppression of gene expression using antisense nucleotide technology, which can be time consuming and in some cases still insufficient to allow assignment of a biological function to a polypeptide encoded by the sequence.

10 The proteome has been defined as the protein complement expressed by a genome. This somewhat restrictive definition implies a static nature of the proteome. In reality the proteome is highly dynamic since the types of expressed proteins, their abundance, state of modification, and subcellular locations are dependent on the physiological state of the cell or tissue. Therefore, the proteome can reflect a cellular state or the external conditions encountered by a cell, and proteome analysis can be viewed as a genome-wide assay to differentiate and study cellular states and to determine the molecular mechanisms that control them. Considering that the proteome of a differentiated cell is estimated to consist of thousands to tens of thousands of different types of proteins, with an estimated dynamic range of expression of at least 5 orders of magnitude, the prospects for proteome analysis appear daunting. However, the availability of DNA databases listing the sequence of every potentially expressed protein combined with rapid advances in technologies capable of identifying the proteins that are actually expressed now make proteomics a realistic proposition. Mass spectrometry

is one of the essential legs on which current proteomics technology stands.

Quantitative proteomics is the systematic analysis of all proteins expressed by a cell or tissue with respect to their quantity and identity. The proteins expressed in a cell, tissue, biological fluid or protein complex at a given time precisely defines the state of the cell or tissue at that time. The quantitative and qualitative differences between protein profiles of the same cell type in different states can be used to understand the transitions between respective states. Traditionally, proteome analysis was performed using a combination of high resolution gel electrophoresis, in particular two-dimensional gel electrophoresis, to separate proteins and mass spectrometry to identify proteins. This approach is sequential and tedious, but more importantly is fundamentally limited in that biologically important classes of proteins are essentially undetectable.

Thus, there exists a need for rapid, efficient, and cost effective methods proteome analysis. The present invention satisfies this need and provides related advantages as well.

SUMMARY OF THE INVENTION

The invention provides methods for identifying a polypeptide. The method can include the steps of determining two or more characteristics associated with the polypeptide, or a fragment thereof, one of the characteristics being mass of a fragment of the polypeptide,

the fragment mass being determined by mass spectrometry; comparing the characteristics associated with the polypeptide to an annotated polypeptide index; and identifying one or more polypeptides in the annotated polypeptide index having the characteristics. The method can further include the steps of determining one or more additional characteristics associated with the polypeptide; comparing the determined characteristics to the annotated polypeptide index; and optionally repeating the steps one or more times, wherein a set of characteristics is determined that identifies a single polypeptide in the annotated polypeptide index. The method can additionally include the step of quantitating the amount of the identified polypeptide in a sample containing the polypeptide. The invention also provides methods for generating a polypeptide identification index such as an annotated polypeptide index.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a schematic diagram of a protein identification strategy based on mass spectrometry (MS) and tandem mass spectrometry (MS/MS) measurements.

Figure 2 shows the steps of a method for comparing and quantitating two polypeptide populations using a polypeptide identification index of Annotated Peptide Tags.

DETAILED DESCRIPTION OF THE INVENTION

The invention provides methods for identifying a polypeptide from a population of polypeptides by determining characteristics associated with a polypeptide, or a peptide

fragment thereof, comparing the determined characteristics to a polypeptide identification index, and identifying one or more polypeptides in the polypeptide identification index having the same characteristics. The methods of the invention are applicable to proteome analysis and allow rapid and efficient identification of one or more polypeptides in a complex sample. The methods are based on generating a polypeptide identification index, which is a database of characteristics associated with a polypeptide. The polypeptide identification index can be used for comparison of characteristics determined to be associated with a polypeptide from a sample for identification of the polypeptide. Furthermore, the methods can be applied not only to identify a polypeptide but also to quantitate the amount of specific proteins in the sample.

The methods of the invention for identifying a polypeptide are applicable to performing quantitative proteome analysis, or comparisons between polypeptide populations that involve both the identification and quantitation of sample polypeptides. Such a quantitative analysis can be conveniently performed in two separate stages, if desired. As a first step, a reference polypeptide index can be generated representative of the samples to be tested, for example, from a species, cell type or tissue type under investigation, as described herein. The second step is the comparison of characteristics associated with an unknown polypeptide with the reference polypeptide index or indices previously generated. A reference polypeptide index is a database of polypeptide identification codes representing the polypeptides of a particular sample, such as a cell, a subcellular fraction

tissue, organ or organism. A polypeptide identification index can be generated that is representative of any number of polypeptides in a sample, including essentially all of the polypeptides potentially expressed in a sample.

- 5 Accordingly, the methods of the invention advantageously allow the determination of polypeptides in a sample that correlates with or defines a particular physiological state of the sample, for example, a disease state. Moreover, once a polypeptide identification index has been generated, the
- 10 index can be used repeatedly to identify one or more polypeptides in a sample, for example, a sample from an individual potentially having a disease.

- For quantitation of a polypeptide in a sample, a polypeptide is compared to a chemically identical molecule
- 15 that is isotopically labeled, for example, with ^{13}C for ^{12}C , deuterium for hydrogen, or ^{18}O for ^{16}O . Any number of differential isotopes can be incorporated so long as there is a sufficient difference in mass to be distinguished by MS, as disclosed herein. Because the molecules are
- 20 chemically identical except for the isotopic difference, the molecules behave physicochemically the same. Furthermore, if desired, more than two samples can be compared if a sufficient number of isotopic labels (e.g., d0, d4, d8, d12) are available such that the multiple samples can be compared
- 25 and distinguished by MS. Quantitation is based on stable isotope dilution. One method to quantitate a sample is to spike a sample with an internal standard that is chemically identical but isotopically different. A standard curve can be generated with dilution of isotope to extrapolate the
- 30 quantity of molecule in a sample. In such a case, the

molecule to be spiked must be identical and therefore the molecules in the sample must be known.

Another convenient method for quantitating polypeptides in a sample is to use a reagent such as ICAT™ (Gygi et al., Nature Biotechnol. 17:994-999 (1999); WO 00/11208). An ICAT™ type reagent, which is described in more detail below, contains an affinity tag, a linker moiety in which one or more stable isotopes can be incorporated, and a reactive group that can covalently couple to an amino acid side chain in a polypeptide such as a cysteine. For quantitation using an ICAT™ type reagent, parallel samples are treated with different isotopic versions of the ICAT™ type reagent. A sample can be labeled and compared to a parallel labeled sample, for example, to normalize to a reference or control sample for quantitation. The use of an ICAT™ type reagent to identify and quantitate polypeptides in a sample is illustrated in Figure 2. Because the peptides labeled with different isotopic versions of the ICAT™ type reagent behave physicochemically the same, the same polypeptides in the two samples will co-purify but still be distinguishable by MS due to the isotopic differences in the ICAT™ type label. Accordingly, the relative amounts of the same polypeptides can be readily compared and quantitated (Gygi et al., *supra*, 1999). Every other scan can be devoted to fragmenting and then recording sequence information about an eluting peptide (MS/MS spectrum). The parent polypeptide that this peptide originated from can be identified by searching a sequence database with the recorded MS/MS spectrum. The procedure thus provides the relative quantitation and identification of the components of protein mixtures in a single analysis.

Such a comparison can be useful for quantitating the expression levels of polypeptides relative to a reference sample, for example, comparing expression levels in a sample from an individual having a disease or suspected of having a disease to a sample from a healthy individual or for forensic purposes.

In addition to being useful for quantitation of polypeptides, an ICAT™ type reagent also functions as a constraint on the complexity of the system, that is, only polypeptides or fragments thereof containing the amino acid reactive with the ICAT™ type reagent will be labeled and characterized if the polypeptides are affinity isolated or compared side-by-side with a differentially isotopically labeled sample (Gygi et al., *supra*, 1999). Accordingly, the use of an ICAT™ type reagent can provide a reduction in complexity of the sample. Furthermore, the ability of a polypeptide or fragment thereof to be labeled with an ICAT™ type reagent, that is, whether the peptide contains the reactive amino acid, is a characteristic associated with the polypeptide useful for identifying the polypeptide in combination with additional characteristics.

An additional advantage of the use of the ICAT™ type reagent is that the identity of polypeptides in a sample need not be known prior to analysis. As described above, isotopic dilution, where an internal standard is spiked into a sample, requires that a chemically identical molecule that is differentially isotopically labeled be spiked into the sample and, therefore, requires that a polypeptide or fragment thereof to be quantitated is known so that a chemically identical isotopically labeled molecule

be added. With an ICAT™ type reagent, no prior knowledge of the exact polypeptides or fragments need be known. Furthermore, there is no need to synthesize a variety of isotopically labeled molecules for characterizing a variety of polypeptides in a sample.

In addition to using a labeling reagent such as an ICAT™ type reagent that incorporates an affinity label, other labeling reagents can be used to differentially isotopically label two different samples containing polypeptides. For example, two chemically identical reagents containing different isotopes can be used to covalently modify two polypeptide samples, where the reagents do not contain an affinity tag. Accordingly, instead of using an affinity isolation step associated with an ICAT™ type tag, other isolation steps, if desired, can be used. Nevertheless, the differentially isotopically labeled polypeptide samples can be compared for quantitative analysis. For example, methylation of polypeptides via esterification with methanol containing d0 (no deuterium) versus d3 (three deuteriums) can be used to differentially isotopically label two polypeptide samples. Similarly, any of the well known methods for modifying side chain amino acids in polypeptides can analogously be used with differentially labeled isotopes such as deuterium for hydrogen, C¹³ for C¹², O¹⁸ for O¹⁶ (see, for example, Glazer et al., Laboratory Techniques in Biochemistry and Molecular Biology: Chemical Modification of Proteins, Chapter 3, pp. 68-120, Elsevier Biomedical Press, New York (1975); Pierce Catalog (1994), Pierce, Rockford IL). Any number of the differential isotopes can be incorporated so long as parallel labeled polypeptides contain a sufficient mass

distinction to be detected by MS. In addition to chemical modification of a polypeptide, as described above, two polypeptide samples can be digested with a protease such as trypsin or the like in the presence of O^{16} - versus O^{18} -
5 labeled H_2O . Since the protease cleavage reaction results in the addition of water to the cleaved peptides, cleavage in the presence of isotopically differentially labeled H_2O can be used to incorporate differential labels into separate polypeptide samples. It is understood that any method
10 useful for incorporating an isotopic label to differentially label two polypeptide samples can be used in methods of the invention, particularly for quantitative methods, so long as the samples to be compared are treated in a chemically similar fashion such that the resulting labeled polypeptides
15 essentially differ only by the differential isotopic label.

Still another method to quantitate a sample is to incubate a sample under conditions that allow metabolic incorporation of isotopes into two samples for comparison by incubating a sample in the presence of an isotope or
20 incubating in media that results in depletion of a naturally occurring isotope (see, for example, Oda et al., Proc. Natl. Acad. Sci. USA 96:6591-6596 (1999)). Such a method is particularly useful for a sample that is conveniently cultured, for example, a microbial sample or a primary
25 culture of cells obtained from an individual. Accordingly, both *in vitro* and *in vivo* methods can be used to differentially isotopically label two samples for comparison and/or quantitation.

The methods of the invention are based on
30 determining characteristics of a polypeptide that allow

identification of the polypeptide based on the determined physicochemical characteristics. The collection of physicochemical characteristics that can function to identify a polypeptide is essentially a "bar code" for the polypeptide, that is, a collection of characteristics sufficient to uniquely identify a polypeptide based on correlating the characteristics with a reference database that functions as a polypeptide identification index. The methods are particularly advantageous for rapid and efficient analysis of complex samples containing many different polypeptides, which would be time consuming and inefficient using other methods. The methods of the invention can thus be applied to analyze complex samples containing numerous different polypeptides and are particularly useful in proteomics applications. Accordingly, the methods of the invention can be advantageously used to identify polypeptides of the proteome. Since the proteome reflects polypeptide expression and post-translational modifications correlated with the metabolic state of the cell, the methods can also be used in diagnostic applications to determine normal or aberrant polypeptide expression associated with a disease. Accordingly, the methods of the invention can be used in clinical applications to diagnose a disease or condition.

The methods of the invention advantageously use constraining parameters that allow the identification of a polypeptide from a complex mixture of different polypeptides. The constraints can be used to simplify the identification of polypeptides. A constraint can be, for example, the inclusion of one or more additional characteristics associated with a polypeptide, the

identification of a subset of polypeptides from a complex mixture, or any type of constraint that can be used to simplify the analysis of a complex mixture of polypeptides. The methods of the invention thus provide more efficient
5 identification of polypeptides in a complex mixture, including large numbers of polypeptides, which is particularly useful for proteome analysis.

The generation and use of a polypeptide identification index provide several advantages. First, the
10 methods can be used with selective isolation of polypeptide fragments containing specific structural features, which can be exploited by tagging with specific chemical reagents. The affinity selection of "tagged" fragments simplifies the polypeptide mixture, rendering it compatible with highly
15 denaturing/solubilizing conditions that can be used for protein isolation and handling. The selective isolation of fragments also constrains database searching. For example, selective cysteine tagging, as disclosed herein, reduces the complexity of the peptide mixture by approximately 10-fold.

20 A second advantage of the invention methods is that they can be readily used in a variety of laboratory settings. For example, mass measurements are absolute and chromatographic parameters can be easily standardized. Therefore, a polypeptide identification index determined by
25 methods of the invention is easily transferable between laboratories, and data generated by different laboratories can be easily compared with a polypeptide identification index generated under similar conditions. This advantage can be further exploited by making the method accessible via
30 a network, for example, through the construction of a Web-

based search tool. A third advantage is that the methods can be performed with a single stage mass analysis, which is fast, simple and sensitive. A fourth advantage is that the methods can be used to accurately measure the ratio of each polypeptide present in a complex polypeptide sample, provided that the samples have been modified with a stable isotope label. Finally, the methods have an essentially unlimited sample capacity, assuring the possibility of analyzing polypeptides of very low abundance, and have a high peak capacity, allowing for the analysis of very complex samples.

In one embodiment of the invention, a polypeptide identification index is generated by determining characteristics associated with a polypeptide, in particular, fragment ion mass measurements by MS/MS optionally including chromatographic steps (LCQ). These mass determinations are not required to be at high accuracy. The accurate mass can be calculated, if desired, and compiled into an index with other characteristics associated with a particular polypeptide. A sufficient number of characteristics are determined to allow identification of a polypeptide in the index. The methods can optionally and advantageously be used with quantitation to provide additional information on the physiological state of a sample. However, in the case of simpler systems, for example, microbial or viral genomes or specimens from an individual containing a smaller number of polypeptides such as spinal fluid, the complexity of polypeptides in a sample can be sufficiently small enough that qualitative analysis of the polypeptides in a sample is sufficient for particular applications. As such, if a qualitative determination of

the expression of a polypeptide in a sample is sufficient to correlate with a particular condition, for example, a disease condition, then the methods of the invention can be applied to a qualitative identification of a polypeptide in a sample.

As used herein, the term "characteristic" when used in reference to a polypeptide refers to a physicochemical property of a polypeptide. Physicochemical properties include physicochemical properties of a parent polypeptide such as molecular mass, amino acid composition, pI and the like, as well as physicochemical properties of a fragment of a polypeptide, including fragment ions, which can be correlated with a polypeptide and are thus considered to be characteristics associated with a parent polypeptide. Physicochemical properties of a polypeptide also include measurable behaviors of a polypeptide that result from its particular physicochemical properties. For example, physicochemical properties include the order of elution on specific chromatographic media under defined conditions, and the position to which a polypeptide migrates in a polyacrylamide gel under defined conditions. The characteristics can be determined empirically or can be predicted based on known information about the polypeptide, for example, sequence information.

As used herein, the term "characteristics associated with a polypeptide" refers to physicochemical properties of a polypeptide and/or any fragment of the polypeptide. As such, the characteristics associated with a polypeptide include specific characteristics of a parent polypeptide as well as characteristics of a fragment of the

parent polypeptide which, because the fragment can be related to the polypeptide, are considered to be characteristics associated with the parent polypeptide. Such characteristics can be used to identify a polypeptide, for example, by comparison with a polypeptide identification index.

As used herein, the term "polypeptide" refers to a peptide or polypeptide of two or more amino acids. A polypeptide can also be modified by naturally occurring modifications such as post-translational modifications, including phosphorylation, lipidation, prenylation, sulfation, hydroxylation, acetylation, addition of carbohydrate, addition of prosthetic groups or cofactors, formation of disulfide bonds, proteolysis, assembly into macromolecular complexes, and the like.

A modification of a polypeptide, particularly ligand polypeptides, can also include non-naturally occurring derivatives, analogues and functional mimetics thereof generated by chemical synthesis, provided that such polypeptide modification displays a similar functional activity compared to the parent polypeptide. For example, derivatives can include chemical modifications of the polypeptide such as alkylation, acylation, carbamylation, iodination, or any modification that derivatizes the polypeptide. Such derivatized molecules include, for example, those molecules in which free amino groups have been derivatized to form amine hydrochlorides, p-toluene sulfonyl groups, carbobenzoxy groups, t-butyloxycarbonyl groups, chloroacetyl groups or formyl groups. Free carboxyl groups can be derivatized to form salts, methyl and ethyl

esters or other types of esters or hydrazides. Free hydroxyl groups can be derivatized to form O-acyl or O-alkyl derivatives. The imidazole nitrogen of histidine can be derivatized to form N-im-benzylhistidine. Also included as
5 derivatives or analogues are those polypeptides which contain one or more naturally occurring amino acid derivatives of the twenty standard amino acids, for example, 4-hydroxyproline, 5-hydroxylysine, 3-methylhistidine, homoserine, ornithine or carboxyglutamate, and can include
10 amino acids that are not linked by peptide bonds.

A particularly useful polypeptide derivative includes modification of sulfhydryl groups, for example, the modification of sulfhydryl groups to attach affinity reagents such as an ICAT™ type reagent. A particularly
15 useful modification of a polypeptide includes modification of polypeptides in a sample with a moiety having a stable isotope. For example, two different polypeptide samples can be separately labeled with moieties that are isotopically distinct, and such differentially labeled samples can be
20 compared. Modification of polypeptides with stable isotopes is particularly useful for quantitating the relative amount of individual polypeptides in a sample.

As used herein, a "fragment" refers to any truncated form, either carboxy-terminal, amino-terminal, or
25 both, of a parent polypeptide. Accordingly, a deletion of a single amino acid from the carboxy- or amino-terminus is considered a fragment of a parent polypeptide. A fragment generally refers to a deletion of amino acids at the N- and/or C-terminus but also includes modifications where a
30 side chain is removed but the peptide bond remains. A

fragment includes a truncated polypeptide that is generated, for example, by polypeptide cleavage using a chemical reagent, enzyme, or energy input. A fragment can result from a sequence-specific or sequence independent cleavage event. Examples of reagents commonly used for cleaving polypeptides include enzymes, for example, proteases, such as thrombin, trypsin, chymotrypsin and the like, and chemicals, such as cyanogen bromide, acid, base, and o-iodobenzoic acid, as disclosed herein. A fragment can also be generated by a mass spectrometry method. Furthermore, a fragment can also result from multiple cleavage events such that a truncated polypeptide resulting from one cleavage event can be further truncated by additional cleavage events.

As used herein, the term "polypeptide identification index" refers to a collection of characteristics associated with a polypeptide sufficient to identify and distinguish other polypeptides in the index. A polypeptide identification index is therefore a collection of polypeptide identification codes for identifying a polypeptide based on characteristics of the polypeptide or a fragment thereof. A polypeptide identification index can be based on deduced characteristics associated with a polypeptide, for example, characteristics predicted based on sequence information such as genomic sequence, cDNA sequence, or EST databases. A polypeptide identification index can also be based on empirically determined characteristics, or a combination of deduced and empirically determined characteristics. An "annotated polypeptide (AP) index" refers to an index comprising at least one empirically determined characteristic for each of the

polypeptides in the index, which can be determined, for example, by the methods disclosed herein. If desired, an AP index can be based on entirely empirically determined characteristics or a combination of deduced and empirically
5 determined characteristics. The use of an annotated polypeptide index is particularly useful for identifying polypeptides modified by post-translational modifications, which can have characteristics unpredictable based on deduction from a sequence database alone.

10 A "polypeptide identification subindex" refers to a subset of a polypeptide identification index that contains less than all of the polypeptide identification codes of the polypeptide identification index. A subindex can contain, for example, five polypeptide identification codes from a
15 polypeptide identification index of ten polypeptide identification codes, which is a subset of the entire index. Identification of a subindex can be useful, for example, for reducing the complexity of a search of a polypeptide identification index, similar to the reduction in complexity
20 that can be applied to a polypeptide sample by the fractionation methods disclosed herein. Accordingly, a search of a subindex can be advantageous in requiring less computational time than required to search an entire index.

As used herein, the term "identification code"
25 refers to a set of characteristics associated with a polypeptide that is sufficient to determine the identity of the polypeptide and distinguish the polypeptide from other polypeptides in a polypeptide identification index. An identification code is essentially an annotated peptide tag,
30 or "bar code," that can be used to identify a polypeptide.

The invention provides a method for identifying a polypeptide. The method includes the steps of determining two or more characteristics associated with a polypeptide or fragment thereof, one of the characteristics being the mass of a fragment of the polypeptide, wherein the fragment mass is determined by mass spectrometry; comparing the characteristics associated with the polypeptide to a polypeptide identification index such as an annotated polypeptide index; and identifying one or more polypeptides in the polypeptide identification index having the determined characteristics. The fragment can be determined at an accuracy in ppm of greater than 1 part per million (ppm) or at even lower accuracy (higher ppm). The method can further include determining one or more additional characteristics associated with the polypeptide and comparing the characteristics determined in each of the steps to the polypeptide identification index. Optionally, the steps of determining one or more additional characteristics associated with the polypeptide and comparing the characteristics determined in each step to the polypeptide identification index can be repeated one or more times, wherein a set of characteristics is determined that identifies a single polypeptide in the polypeptide identification index. The method can further include quantitating the amount of polypeptide in a sample. Furthermore, the methods can be used to measure the relative abundance in two or more different populations of polypeptides, that is, polypeptide mixtures, for example, populations of polypeptides in different samples.

The methods of the invention for identifying a polypeptide include determining characteristics associated

with the polypeptide, or a fragment of the polypeptide. Characteristics associated with a polypeptide that are useful for identifying a polypeptide are those characteristics that can be reproducibly determined.

- 5 Physicochemical properties of a polypeptide or fragment include, for example, atomic mass, amino acid composition, partial amino acid sequence, apparent molecular weight, pI, and order of elution on specific chromatographic media under defined conditions. Such characteristics determined to be
10 associated with a polypeptide are used for the identification of the polypeptide. Methods for determining characteristics associated with a polypeptide are described in more detail below.

- One of the characteristics particularly useful in
15 methods of the invention is the mass of a polypeptide or fragment or fragments thereof. A fragment of a polypeptide can be generated prior to or during the process of mass determination by mass spectrometry. A polypeptide fragment mass can therefore be the mass of a fragment of a
20 polypeptide generated during polypeptide sample preparation, or can be the mass of fragment generated by a polypeptide cleavage that occurred during mass spectrometry.

- In the methods of the invention, the mass of a polypeptide fragment is determined by mass spectrometry.
25 The methods of the invention allow the identification of a polypeptide without the need for sequencing the polypeptide or fragment thereof. A polypeptide fragment mass can be determined using a variety of mass spectrometry methods known in the art, as described herein.

A variety of mass spectrometry systems can be employed in the methods of the invention for identifying a polypeptide. Mass analyzers with high mass accuracy, high sensitivity and high resolution include, but are not limited to, matrix-assisted laser desorption time-of-flight (MALDI-TOF) mass spectrometers, ESI-TOF mass spectrometers and Fourier transform ion cyclotron mass analyzers (FT-ICR-MS). Other modes of MS include an electrospray process with MS and ion trap. In ion trap MS, fragments are ionized by electrospray or MALDI and then put into an ion trap. Trapped ions can then be separately analyzed by MS upon selective release from the ion trap. Fragments can also be generated in the ion trap and analyzed. The ICAT™ type reagent labeled polypeptides that can be used in the methods of the invention can be analyzed, for example, by single stage mass spectrometry with a MALDI-TOF or ESI-TOF system.

If desired, different MS analysis can be applied for generating a polypeptide identification index than for determining characteristics of an unknown polypeptide. For example, LC-MC-MS can be used for collecting data for the identification index and LC-EI-TOF can be used for measurement of characteristics of an unknown polypeptide. It is understood that any MS methods and any combination of MS methods can be used so long as the samples are treated in a substantially similar manner and so long as the MS methods are compatible for comparison of masses determined by the different methods.

The methods of the invention can involve a polypeptide separation step followed by a mass analysis step. Polypeptide separation and mass analysis steps can be

performed independently or can be coupled in an "on line" analysis method. Various modes of polypeptide separation techniques can be coupled to a mass analyser. For example, polypeptides can be separated by chromatography using
5 microcapillary HPLC, by solid phase extraction-capillary electrophoresis systems that can be coupled to a mass analyzer, or by gel electrophoresis methods. A specific example of a coupled polypeptide separation and mass
10 ESI-MS/MS system that is applied with dynamic exclusion on an ion trap MS.

Different types of mass spectrometry can be used for different applications of the methods of the invention.
15 For certain applications, such as mass determination of a polypeptide fragment for generating a polypeptide identification index, a method that provides high accuracy, such as an accuracy of less than 1 part per million. However, the methods of the invention are advantageous in
20 that MS of lower accuracy, that is higher ppm resolution, can be conveniently used without the need for more expensive instrumentation required for higher accuracy determinations. For applications that involve high throughput analysis of a population of polypeptides, a lower accuracy mass
25 determination can be sufficient. Lower accuracy mass determinations generally provide higher sample throughput because less time is required to make a mass determination.

The methods of the invention involving mass determinations can be conveniently performed at lower
30 accuracy. For example, high mass accuracy instruments such as FTMS or FTICR MS can be used to determine accuracy at 0.2

ppm (Goodlett et al. et al., Anal. Chem. 72:1918-1924 (2000)). The use of very high mass accuracy such as 0.1 ppm acts as a constraint. However, the methods of the invention are advantageous in that several characteristics associated with a polypeptide can be determined. When combined with additional characteristics, the masses can be determined at lower accuracy, that is higher ppm. Determination of mass at lower accuracy allows the use of less expensive MS instruments which are more widely available than FTMS. The mass determinations can be determined at an accuracy in ppm of 1 part per million (ppm) or greater than 1 ppm, and can be determined at an accuracy in ppm of 2.5 ppm or greater, of about 5 ppm or greater, about 10 ppm or greater, about 50 ppm or greater, about 100 ppm or greater, about 200 ppm or greater, about 500 ppm or greater, or even about 1000 ppm or greater, sequentially each of which requires less accuracy of the MS instrument. The methods of the invention advantageously allow the use of lower accuracy MS analysis in combination with other physicochemical characteristics, as disclosed herein to identify a polypeptide in a sample. The accuracy of the MS measurement for a particular application can be readily determined by one skilled in the art, for example, depending on the complexity of the sample and/or index to be used.

The methods of the invention for identifying a polypeptide can involve determining the mass of a polypeptide fragment at an accuracy of greater than 1 part per million. Therefore, the method does not require a MS method having high accuracy. Accordingly, a lower-cost MS system can be employed in the methods of identifying a polypeptide. The adaptation of any mass spectrometry to a

high throughput format, such as 96-well plate or 384 spot plate format, or to an autoinjection system that allows unattended operation, is advantageous for increasing sample throughput.

5 An overview of the strategy of a protein identification method is shown in Figure 1. Polypeptides are optionally fractionated from a mixture, for example, using polyacrylamide gel electrophoresis, and the polypeptides can further be fragmented into peptides. The
10 peptides can further be optionally fractionated by chromatography. A chromatographic fraction, or bin (indicated by "*" in Figure 1), is subjected to MS. Traditionally, an ion or dominant ions are selected in a collision cell for collision-induced dissociation (CID).
15 Selection of a single ion is depicted in Q1 of Figure 1. An ion is selected and then fragmented, as shown in Q3 of Figure 1. As shown in the bottom of Figure 1, the fragment masses can be combined with any number of additional characteristics and compared to a protein identification
20 index, for example, a sequence database, and the polypeptide is identified based on those determined characteristics.

A set of determined characteristics associated with a polypeptide are compared to the characteristics associated with a polypeptide in a polypeptide
25 identification index. A polypeptide identification index is a collection of characteristics associated with individual polypeptides that uniquely identify and distinguish the polypeptides from other polypeptides annotated in the index. By comparing the set of determined characteristics
30 associated with a polypeptide to a polypeptide

identification index, one or more polypeptides in the polypeptide identification index that share the same characteristics can be identified. If more than one polypeptide is determined to have the same characteristics, 5 additional constraints can be included, for example, the determination of one or more additional characteristics. A polypeptide identification index can be based on deduced characteristics of a polypeptide, for example, one or more characteristics deduced from genetic sequence databases, or 10 can be determined empirically, as with the annotated peptide tag index described herein.

One exemplary method of generating an annotated peptide index is to: Harvest Proteins; Label Proteins with an Isotope Coded Affinity Tag (ICAT™) Reagent; Fractionate 15 Proteins by Molecular Weight; Digest Proteins to Peptides (e.g. using Trypsin); Separate Peptides by Ion Exchange; Purify each Ion Exchange Fraction by Affinity chromatography; Analyze each Affinity chromatography Fraction by LC/MS/MS (or CE/MS/MS); Identify all Expressed 20 Proteins via Database Search of individual MS/MS Peptide Spectra; Generate a Database of Annotated Peptide Tags that constitute a unique barcode for an individual; Peptide based on measured Physicochemical properties and thus the Parent Protein of that Peptide. It is understood that the above- 25 described method, combinations of these steps, modifications thereof, or any methods suitable to allow the determination of characteristics associated with a polypeptide can be used to generate a polypeptide identification index containing at least one empirically determined characteristic, as 30 described herein.

The methods of the invention can further include determining one or more additional characteristics associated with the polypeptide for comparison with a polypeptide identification index. The process of
5 determining one or more additional characteristics associated with a polypeptide followed by comparing with a polypeptide identification index can be repeated until a single polypeptide is uniquely identified from the polypeptide identification index. Accordingly, if
10 additional constraints are applicable, they can be included to identify a polypeptide by comparison to a polypeptide identification index.

The number of characteristics sufficient to identify of a polypeptide can be readily determined by one
15 skilled in the art by comparing the determined set of characteristics with the polypeptide identification index. The identification of a single polypeptide in a polypeptide identification index refers to determining a set of characteristics that are sufficient to distinguish the
20 polypeptide from another polypeptide in the polypeptide identification index. For example, if two determined characteristics match a single polypeptide in a polypeptide identification index, then the two characteristics are sufficient to identify a single polypeptide. Similarly, for
25 a different polypeptide, three determined characteristics can be required to uniquely identify a polypeptide in the index. Accordingly, based on the characteristics determined for a polypeptide, a comparison is made to a polypeptide identification index. If a single polypeptide is
30 identified, then a sufficient number of characteristics have been determined. If more than one polypeptide is

identified, then one or more additional characteristics can be determined until a single polypeptide uniquely matches the determined characteristics, thereby allowing identification of the polypeptide. Therefore, one skilled in the art can readily determine if a sufficient number of characteristics, based on comparison to a particular polypeptide identification index, have been determined for a polypeptide to allow identification of a unique polypeptide in the polypeptide identification index.

10 The methods of the invention are advantageously based on the inclusion of selected constraints that allow more efficient identification of a polypeptide, particularly in complex samples containing numerous different polypeptides. The methods can also be advantageously used to identify multiple polypeptides from a complex sample. Accordingly, rather than determining a large number of characteristics associated with different polypeptides, the methods can be performed in an iterative manner, if desired, with the inclusion of additional constraints as needed to identify a single polypeptide in a polypeptide identification index.

For example, polypeptides that are homologous generally have segments of high sequence identity. Such polypeptides can arise, for example, from polypeptides having similar function, splice variants of the same nucleic acid, and the like. Polypeptides having segments of high sequence identity can have in common several physicochemical characteristics, particularly in association with homologous fragments of the polypeptide. Polypeptides sharing a high degree of similarity can therefore have a similar or

identical set of associated characteristics. For such similar polypeptides, a given set of characteristics sufficient to distinguish two dissimilar polypeptides can be insufficient for the identification of a single polypeptide in a polypeptide identification index when the polypeptides have regions of similarity. In such a case, one or more additional characteristics associated with the polypeptide can be determined, and the determination of additional characteristics can be repeated until the subject polypeptide can be distinguished from each other polypeptide in a polypeptide identification index. The methods of determining a set of characteristics associated with a polypeptide, comparing with a polypeptide identification index, and determining additional characteristics until a single polypeptide in a polypeptide identification index is identified can be applied to one or more polypeptides.

Thus, additional constraints, as needed to identify a polypeptide, can be considered. For example, if more than one polypeptide in a polypeptide identification index has a given set of characteristics, the identification of selected polypeptides of the polypeptide identification index, that is, a subset of polypeptides in the index or a subindex of the index, functions essentially as a constraint. Accordingly, a subsequent comparison to the polypeptide identification index can be made to the subindex, which can reduce the calculation time and provide a more efficient comparison, if desired. An additional constraint can then be considered, for example, an additional characteristic, and compared to the subindex, which can result in a reduction in the number of polypeptides having all of the determined characteristics.

Such steps can be optionally repeated until a single polypeptide in the polypeptide identification index is identified. Such an approach is advantageous when determining the identity of multiple polypeptides because
5 only those characteristics sufficient to identify a polypeptide need be determined. The methods can thus readily accommodate the determination of the identity of a variety of polypeptides and the complexities associated with proteomics analysis without wasting resources on unnecessary
10 data acquisition.

The methods of the invention for generating a polypeptide identification index involve determining, for two or more polypeptides, a set of characteristics that can be used to identify the polypeptide. A set of
15 characteristics that uniquely identify a polypeptide in a polypeptide identification index define a polypeptide identification code or "bar code" for the polypeptide. A polypeptide identification index can contain a variety of characteristics associated with an indexed polypeptide.
20 Polypeptide characteristics contained in a polypeptide identification index can include polypeptide mass, amino acid composition, partial amino acid composition, for example, the presence of a particular amino acid, pI, order of elution on specific chromatographic media, and one or
25 more polypeptide fragment masses. A polypeptide identification index can additionally include amino acid sequence, references to related polypeptides, database entries or literature, as well as other information relevant to the identification of a polypeptide. The user will know
30 what types of information are useful for a polypeptide index and can include any physicochemical property or information

relating to a polypeptide. A polypeptide identification index containing a large number of identification codes for a variety of polypeptides is particularly useful for identifying polypeptides in complex samples.

5 The methods of the invention directed to identifying a polypeptide are based on comparing characteristics determined for a polypeptide with a polypeptide identification index. A polypeptide identification index can be a commercially or publicly
10 available database such as GenBank (www.ncbi.nlm.nih.gov/GenBank), in which one or more characteristics of a polypeptide are predicted, for example, amino acid composition, mass of a polypeptide or fragment thereof, and the like. In addition, a polypeptide
15 identification index can be based on empirically determined characteristics determined by the methods described herein. In addition, a polypeptide identification index can be a combination of predicted and empirically determined characteristics, for example, like the annotated polypeptide
20 (AP) index disclosed herein, also referenced as an annotated peptide tag (APT) index.

 The set of characteristics associated with a polypeptide can be determined experimentally using a variety of methods. An exemplary method for polypeptide
25 identification and/or determining characteristics for generating a polypeptide identification index is shown in Figure 1 and described below. The method is useful for defining a polypeptide identification code because the method involves a series of steps, which allow the
30 determination of characteristics associated with a

polypeptide, the final step being mass determination of a polypeptide or fragment. The method can include: (i) polypeptide sample preparation; (ii) polypeptide tagging; (iii) optional polypeptide fractionation; (iv) polypeptide fragmentation.; (v) polypeptide fragment separation; (vi) affinity isolation of tagged polypeptide fragments; (vii) high resolution polypeptide fragment separation; (viii) database searching; and (ix) polypeptide identification index construction (see Example I).

10 For polypeptide sample preparation, polypeptide samples for which quantitative proteome analysis is to be performed are isolated from the respective sources using standard protocols for maintaining the solubility of all the polypeptides. Polypeptide samples and preparation of
15 polypeptide samples are discussed in detail below.

For polypeptide tagging, the polypeptides in the sample can be denatured, optionally reduced, and a chemically reactive group of the polypeptides is covalently derivatized with a chemical modification reagent. An
20 exemplary reactive group is a sulfhydryl group that represents a side chain of a reduced cysteine residue, which can be derivatized by a reagent such as ICAT™ (Gygi et al., Nature Biotechnol. 17:994-999 (1999)) or IDEnT reagent (Goodlett et al., Anal. Chem. 72:1918-1924 (2000)). Other
25 useful reactive groups include amino or carboxyl groups of polypeptides or specific post-translational modifications, including phosphate, carbohydrate or lipid. Any chemical reaction with specificity for a chemical group in the polypeptide can be applied in this step. The ICAT™ type and

IDeNT reagents and method of use are described in more detail below.

For optional polypeptide fractionation, the mixture of tagged polypeptides can be fractionated using any polypeptide separation procedure. A fractionation procedure useful in methods of the invention is reproducible, allows polypeptides to remain soluble, and has a high sample and peak capacity. Any optional fractionation technique can be performed to enrich for low abundance proteins and/or to reduce the complexity of the mixture, while the relative quantities are maintained. Exemplary fractionation methods include, for example, sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), chromatographic methods such as size exclusion, ion exchange, and the like, as disclosed herein. Polypeptide fractionation methods are described in more detail below.

For polypeptide fragmentation, the polypeptides in the sample mixture, or the polypeptides contained in each fraction if optional sample fractionation is employed, can be subjected to sequence specific cleavage, such as cleavage by trypsin. The use of sequence specific cleavage can be particularly useful because the termini of peptides cleaved by a sequence specific method can act as a constraint. However, it is understood that the cleavage method used to generate fragments need not be sequence specific, if desired. Methods useful for cleaving polypeptides in a sequence specific manner are described in more detail below.

For polypeptide fragment separation, the resulting polypeptide fragment mixtures can optionally be subjected to

a first dimension peptide separation. Separation methods having a high sample capacity, at least moderate resolving power and highly reproducible separation patterns are useful in this step. Examples of first dimension separation
5 methods include anion and cation ion exchange chromatographies. Chromatographic methods are described in more detail below. Although polypeptide fragment separation can optionally be performed, the methods can be advantageously used such that the characteristics of peptide
10 fragments are measured in "bulk," that is, the methods do not require peptide fragment purification to homogeneity.

For affinity isolation of tagged polypeptide fragments, polypeptide fragments can be isolated from each chromatographic fraction using an affinity reagent that
15 binds to the polypeptide tag. For example, polypeptide fragments tagged with the ICAT™ type reagent exemplified herein can be isolated using avidin or streptavidin affinity chromatography. An example of a useful affinity medium for isolation of ICAT™ type labeled polypeptide fragments is
20 monomeric avidin immobilized on polymer beads. If ICAT™ type reagents with affinity tags different from biotin are used, corresponding affinity media that binds the affinity tag is used.

For high resolution polypeptide fragment
25 separation, liquid chromatography ESI-MS/MS can be used. The polypeptide fragment mixtures eluted from the affinity chromatography columns can be individually analyzed by automated LC-MS/MS using capillary reversed phase chromatography as the separation method (Yates et al.,
30 Methods Mol. Biol. 112:553-569 (1999)) and data dependent

CID with dynamic exclusion (Goodlett, et al., *supra*, 2000) as the mass spectrometric method.

For database searching, the sequence of polypeptide fragments for which suitable CID spectra were
5 obtained are determined by searching a sequence database from the species under investigation. A sequence database search program such as SEQUEST® (Eng, J. et al., J. Am. Soc. Mass. Spectrom. 5:976-989, (1994)) or a program with similar capabilities can be advantageously used to search a
10 database.

For polypeptide identification index construction, the sequences of all the peptides that have been identified by the procedure described above can be entered in a database and annotated with characteristics that were
15 generated during the above-described steps. These attributes can include, for example, partial amino acid composition, the approximate molecular mass of the parent polypeptide, which can be determined, for example, by the optional fractionation step, the order of elution from a
20 first chromatography step, the order of elution time from a second chromatography step, and the like.

Collectively, a sufficient number of characteristics can be determined that distinguish each polypeptide fragment in a polypeptide identification index.
25 The collection of characteristics that uniquely identify a polypeptide represent a "bar code" or polypeptide identification code. Characteristics associated with an unknown polypeptide can be subsequently determined and compared to a previously generated polypeptide

identification index. Alternatively, a polypeptide identification index can be determined along with the unknown polypeptide. However, the accumulation of information related to characteristics associated with a polypeptide and collection in an index is convenient for minimizing the experimental steps needed at the time of analyzing a sample. Therefore, a polypeptide identification code that is determined for a fragment of a polypeptide generated in a subsequent experiments can be used to identify a polypeptide in a sample by correlating the polypeptide identification code newly generated for an unknown polypeptide with the polypeptide identification code to identify the unknown polypeptide.

For the identification of a polypeptide by comparison with a polypeptide identification index, a set of characteristics associated with a polypeptide can be determined generally as described above, or can be determined using an equivalent, modified or abbreviated method, or any method that allows for the determination of characteristics associated with a polypeptide. The number of characteristics sufficient to uniquely identify a polypeptide can be readily determined by those skilled in the art. The methods will generally include the identification of 2 or more characteristics, and can include 3 or more, 4 or more, 5 or more, 6 or more, 7 or more, 8 or more, 9 or more, or even 10 or more characteristics, or any number of characteristics so long as a sufficient number of characteristics are determined that distinguish each of the polypeptides in the index.

In generating a polypeptide identification index, characteristics associated with a polypeptide can be used to obtain the polypeptide sequence by searching a sequence database. For example, a partial amino acid sequence of a polypeptide or fragment optionally determined by mass spectrometry can be readily used to search a polypeptide or translated nucleic acid sequence database to identify a name or sequence identification number, such as an accession number, that uniquely describes a polypeptide. A polypeptide identification index can therefore contain polypeptide characteristics such as a common name, a numeric or alphanumeric identification code from a publicly available database, or any other identifying code selected for identifying a polypeptide identification code in a polypeptide identification index.

To obtain sequence information from polypeptides that do not have a parent polypeptide or nucleic acid sequence in a database or that contain an unexpected post-translational modification that presents identification, *de novo* sequencing can be performed. Identified amino acid sequence can be used to search a polypeptide or nucleic acid sequence database as described above. *De novo* sequencing can be performed using a variety of methods. A particularly useful method of *de novo* sequencing involves using a MS dataset generated for polypeptide identification.

It is understood that, although sequence information regarding a polypeptide or portion thereof, for example, determined by a method such as CID, can be included as a characteristic in a polypeptide identification index, the methods of the invention obviate the need to sequence an

unknown polypeptide in order to identify it, although sequence information can be included in generating a polypeptide identification index, if desired. Accordingly, a polypeptide identification index can contain information
5 on characteristics associated with a polypeptide that is additional to those characteristics sufficient to identify a polypeptide, for example, sequence information. By accumulating information regarding characteristics associated with a polypeptide in an index, the identity of a
10 polypeptide can be readily determined in the absence of obtaining sequence information on the unknown polypeptide.

A chromatographic separation can be used to determine a characteristic of a polypeptide because the physicochemical properties of a polypeptide are reflected in
15 the behavior of the polypeptide on chromatographic media. For example, a highly charged polypeptide will be eluted from an anion or cation exchange column under specific pH and/or salt conditions that differ from the pH and/or salt conditions under which an uncharged or oppositely charged
20 polypeptide will elute. Therefore, a characteristic associated with a polypeptide can be the particular pH and/or salt condition under which the polypeptide is eluted from a chromatographic column. Similarly, conditions under which a polypeptide elutes from any type of chromatographic
25 column can be determined. An order of elution or buffer condition at which a polypeptide is eluted from a column can be assigned a value to be annotated in a polypeptide index or to be used for comparing with corresponding values in a polypeptide index. A value can be, for example, relative
30 position in an elution profile under defined conditions, a time of elution under a given set of conditions and flow

rate, the relative time or order of elution in relation to an external standard fraction number, salt concentration, pH, or any parameter that describes the behavior of a polypeptide on a particular chromatography column that can
5 be reproducibly determined. Alternative methods include gel electrophoresis, for example, isoelectric focusing (IEF) or other analytical electrophoretic methods. Methods for fractionating polypeptides are well known to those skilled in the art (Scopes, Protein Purification: Principles and
10 Practice, 3rd ed., Springer Verlag, New York (1993)).

Protein fractionation steps are useful in the methods of the invention for both reducing the complexity of a polypeptide sample prior to mass analysis of a polypeptide or fragment thereof and for determining characteristics
15 associated with a polypeptide. Any of the well known fractionation steps, in addition to chromatographic fractionation described above, can be used to reduce the complexity of the sample and/or serve as a determined characteristic associated with a polypeptide. Exemplary
20 fractionation steps include salt precipitation such as ammonium sulfate or precipitation with chemicals such as polyethylene glycol or polyethyleneimine, subcellular fractionation, tissue fractionation, immunoprecipitation, and the like (see Scopes, *supra*, 1993). A fractionation
25 step can be used to reduce the complexity of a polypeptide population. For example, complexity reduction can be used in the isolation of a polypeptide subpopulation containing polypeptides tagged on a particular amino acid.
Furthermore, other fractionation steps such as subcellular
30 fractionation can also be applied to reduce the complexity of a sample and/or provide a characteristic useful for

identifying a polypeptide. The fractionation steps can potentially provide biologically important information on the polypeptide, for example, whether the polypeptide is located in an organelle or is a nuclear protein, a membrane protein, and/or part of a signaling complex, and the like. Any fractionation step that advantageously reduces polypeptide population complexity can be applied in the methods of the invention.

A polypeptide fractionation step is useful in the methods of the invention for determining a characteristic associated with a polypeptide. For example, a protein fractionation method based on molecular weight can be used to determine a polypeptide molecular weight. Methods such as SDS-PAGE, commercially available gel elution or preparative cell systems (BIO-RAD), and size exclusion chromatography can be used to determine the apparent molecular weight of a polypeptide or fragment. Polypeptide and/or fragment molecular weight is a characteristic that can be included in a polypeptide identification index.

The particular set of characteristics determined for a polypeptide in generating a polypeptide identification index or for identifying a polypeptide can be selected by the user and will depend on the polypeptide sample, the methods used to prepare the polypeptide sample, the method of mass spectrometry employed and the preferences of the user. The characteristics of a polypeptide can be obtained in any temporal order. For example, polypeptide characteristics can be collected in an order that provides time efficiency or convenience or can be collected as

dictated by a particular method selected for sample processing.

In generating a polypeptide index, sequence information, for example, determined by CID, as well as
5 other characteristics of a polypeptide can be used, and the sequence information is particularly useful for correlating other characteristics of a polypeptide with a particular sequence to identify the polypeptide. However, the methods are advantageous in that, once a polypeptide identification
10 index has been generated, obtaining sequence information on a polypeptide is not required. Instead, other characteristics sufficient to identify a polypeptide can be determined, for example, masses and/or ratios between peptides as well as other characteristics, and compared to a
15 polypeptide identification index, which itself can include sequence information, thereby eliminating the need to sequence a polypeptide in order to identify it.

The methods of the invention for generating a
20 polypeptide identification index involve determining a set of characteristics associated with a first and second polypeptide in which the determined characteristics are sufficient to distinguish the first and second polypeptides. Characteristics that are sufficient to distinguish the first
25 and second polypeptides refer to a set of characteristics that can be uniquely attributed to a polypeptide so that the polypeptide identity can be determined unambiguously. In a case in which set of characteristics is shared by one or more polypeptides, an additional characteristic that allows
30 a polypeptide to be distinguished from another polypeptide is determined. Thus, the polypeptides represented in a

polypeptide identification index can be distinguished from each other by the set of characteristics that identify each polypeptide.

The methods of the invention for identifying a polypeptide can be applied to a population of polypeptides in which two or more polypeptides are identified and can be conveniently used to identify multiple polypeptides in a sample, if desired. Therefore, the method can be applied to a simple or complex polypeptide sample. A simple polypeptide sample can be, for example, a purified polypeptide sample containing one to several polypeptides. A complex sample can be, for example, a cell lysate or fraction containing a few to several hundred polypeptides or even thousands or tens of thousands of polypeptides. Using the methods described herein, the determination of polypeptide characteristics can require the collection of experimental data resulting from a series of steps, such as, for example, a series of chromatographic separations.

An exemplary process useful for organizing data obtained during analysis of complex polypeptide samples involves parceling information into theoretical "bins". For example, an ICAT™ type reagent-labeled mixture of polypeptides can be separated by size into a particular number of bins, which can be fractions eluting from chromatography column, such as size exclusion, ion exchange, and the like, or segments of an SDS polyacrylamide gel. The polypeptides in each bin can be fragmented by a sequence specific cleavage method. Alternatively, analysis of polypeptides in a sample can be performed without fractionating the polypeptides so long as there has been a

sufficient reduction in complexity of the sample to allow the identification of the polypeptide without fractionation. The peptide mixture, which has been fractionated into bins, can be further fractionated by various methods, including, for example, ion exchange chromatography, affinity chromatography such as is used with the isolation of ICAT™ labeled peptides, or reverse phase liquid chromatography. Each bin of peptides can then be further binned by ion exchange chromatography and once again divided further into a particular number of bins. Each of these bins can be further separated by reverse phase chromatography and divided further into a particular number of bins, each of which can be analyzed by mass spectrometry. Hence each polypeptide analyzed by such a method will have five associated characteristics that can be represented, for example, as a 5-digit polypeptide identification code or "bar code" based on cysteine content, size, charge, hydrophobicity, and mass.

The methods of the invention for indexing characteristics associated with a large number of polypeptides use an amount of computer memory that is quadratic in sequence length. An advanced data structure such as, for example, suffix trees, can be used to reduce the requirements of computer memory (Gusfield, Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology, Cambridge University Press (1997)). Suffix trees are a compact data representation for all suffixes in a database of sequences. In their pure form, they can be constructed in linear time and stored in linear, instead of quadratic memory. Various modifications of suffix trees and traversal algorithms can be used to

optimize computation time and use of computer memory associated with searching a polypeptide identification index.

A set of determined characteristics associated with a polypeptide are compared to a polypeptide identification index. Various search algorithms can be employed for matching values assigned to determined characteristics with annotated values in the index. A useful strategy for increasing the efficiency of database searching is the narrowing or "constraining" of the database. The term "constrain" when used in reference to a polypeptide identification index refers to a limitation that is applied to a polypeptide identification index in order to obtain a subindex containing a fraction of polypeptide identification codes corresponding to polypeptides having characteristics that match one or characteristics of a polypeptide to be identified. A subindex can be generated when a group of polypeptides having a common characteristic is selected out of a polypeptide identification index or when a particular characteristic contained in a polypeptide identification code is used to omit one or more polypeptides from an index. A common characteristic can be a definite physicochemical characteristic such as a partial amino acid sequence or any other determined characteristic assigned a range of values. For example, a mass of a polypeptide fragment expressed as a range of values that account for the error in mass determination can serve as a constraint for selecting a subset of polypeptides or fragments of a particular mass.

One characteristic associated with a polypeptide that can be used to constrain a database is partial amino acid composition. The partial amino acid composition of a polypeptide includes the identification of a single amino acid present in a particular polypeptide or fragment thereof. A partial amino acid sequence can be obtained, for example, by treating a polypeptide or fragment thereof with a reagent that results in the generation of a polypeptide or fragment that contains one or more defined amino acids. For example, a sequence specific polypeptide cleavage method will produce fragments with one or more known amino acid residues at the fragment carboxy- or amino-terminus. However, it is not necessary to know if a specific amino acid residue is located at the fragment carboxy- or amino-terminus of a polypeptide. Accordingly, cleavage of a polypeptide with a sequence specific protease indicates the presence of the corresponding amino acid and/or sequence in the polypeptide or peptide fragment thereof. Similarly, a reagent can be used to specifically modify or label one or more specific amino acid residues of a polypeptide or fragment. A polypeptide or fragment that contains such a modification or label will be known to contain a specific amino acid. Partial amino acid composition is a characteristic associated with a polypeptide that can be useful for constraining a polypeptide identification index to generate a polypeptide identification subindex.

The comparison of a set of determined characteristics with a polypeptide identification index can therefore involve a series of searches constrained by a determined characteristic of a polypeptide. For example, an initial search of parent polypeptide or fragment mass can be

performed, resulting in the generation of a polypeptide identification subindex containing polypeptide and fragment mass values that are similar to, that is, within the range of instrument error, the polypeptide or fragment thereof to be identified. A second characteristic to be searched against the generated polypeptide identification subindex, such as the presence of a cysteine residue in the polypeptide to be identified, provides a further constraint and can be used to generate a further polypeptide identification subindex.

The determined mass of a polypeptide or fragment is a characteristic that can be advantageously used to constrain such a database search to increase the efficiency of searching a large database. For example, tandem MS spectra can be analyzed using software such as SEQUEST®, which generates a list of peptides in a database that match the molecular mass of the unknown peptide on which CID was carried out and then compared the observed CID spectrum of the unknown with that for all possible isobars (Eng, J. et al., J. Am. Soc. Mass. Spectrom. 5:976-989, (1994)). Therefore, the set of peptides having a molecular mass similar to the polypeptide fragment being analyzed generated by this type of search provides a subset of possible parent polypeptides represented by the polypeptide fragment. The subset can then be searched using, for example, a partial amino acid composition, to identify the parent polypeptide. Those skilled in the art will know or can readily determine appropriate correlation score parameters for a particular search using software applications such as SEQUEST®.

The method of comparing two or more polypeptide populations employs a method for quantitatively distinguishing the two polypeptide populations, such as the method described herein using an ICAT™ type reagent and is illustrated in Figure 2. Two or several chemically identical but differentially isotopically labeled ICAT™ type reagents can be used at this step. Therefore, although Figure 2 depicts two samples, multiple samples can be compared using the methods described herein. The samples depicted in Figure 2 contain polypeptide populations harvested from the same sample type that differ from each other in growth condition. Exemplary differential growth conditions can include growth under different metabolic conditions or cells at different metabolic states, comparison of a normal and disease sample such as a tumor sample, comparison of untreated versus cells treated with a pharmacological agent, and the like.

As shown in Figure 2, the samples are independently labeled using the ICAT™ type reagent, combined, and characteristics of the polypeptides and corresponding fragments are determined, as described herein. Polypeptides and fragments generated during this process can be analyzed using single stage mass spectrometry, rather than by MS/MS, if desired, to increase sample throughput and sensitivity (Goodlett et al., *supra*, 2000). The characteristics determined for polypeptides and fragments are used to determine polypeptide identities, as described herein. Subsequently, the mass spectra can be examined for pairs of peptide ions that co-fractionated throughout the process and that have a mass difference that precisely corresponds to the mass difference encoded in the ICAT™ type

reagent. The relative signal intensities of the two peaks indicate the relative abundance of the fragment polypeptides and therefore indicate the relative abundance of the corresponding parent polypeptide initially present in the sample. Therefore, a method for comparing the polypeptides contained in two polypeptide samples can involve the generation of two reference polypeptide indices that contain, for each polypeptide identified, a quantitative determination of polypeptide amount in addition to a polypeptide identification code.

An alternative method for comparing two or more polypeptide populations is the comparison of one or more polypeptide samples to a previously determined polypeptide reference index. A set of characteristics of one or more polypeptides in a polypeptide sample can be identified and compared to a reference polypeptide identification index to determine the identities of one or more polypeptides and comparative quantities of the identified polypeptides. If desired, an unknown sample can be compared to a reference sample using the above-described quantitative methods to determine relative expression levels of the polypeptides. A reference sample can be, for example, a sample from a healthy individual or a sample from a control condition useful for comparing to the physiological state of another sample such as a disease sample.

A polypeptide identification index that contains quantitative determinations of polypeptide amount is considered to be a "polypeptide profile" of the particular sample used to generate the index. A polypeptide profile, as used herein, is a set of polypeptide identification codes

that includes polypeptide amount, generated for a specific sample.

A polypeptide profile is useful in methods of proteomics because such a profile can be used to distinguish
5 between different conditions or states of cells, tissues, organs, and organisms. The polypeptides expressed by a cell or tissue at a particular time can be used to define the state of the cell or tissue at the time of measurement. Therefore, quantitative and qualitative differences between
10 the polypeptide profiles of the same cell type in different states can be used to diagnose the respective states. Examples for such comparisons include normal versus tumor cells, cells at different metabolic states and untreated cell versus cells treated with specific pharmacological
15 agents. The differences between two polypeptide profiles can be described as a "differential polypeptide profile".

A differential polypeptide profile is useful for analyzing quantitative changes in the polypeptides contained in samples derived from different cell types such as, for
20 example, cancerous and normal cells, stimulated and unstimulated cells, or from different tissue samples of clinical interest.

The methods of the invention for generating differential polypeptide profiles are applicable to the
25 analysis of changes in the polypeptide profiles in samples such as body fluids. A differential polypeptide profile is determined by comparing the polypeptide profile of two specimens, for example, a normal to disease-related polypeptide profile. For example, a polypeptide profile

representative of a normal specimen state can be generated and compared to a specimen suspected to be in an abnormal or disease state. Alternatively, a reference polypeptide profile representative of a disease state can be compared
5 with a specimen from an individual having or suspected of having a particular disease state. A reference polypeptide profile representative of a normal or disease state can be determined using a specimen from a particular individual or a population of individuals.

10 If desired, analysis can be performed on a population rather than an individual, particularly a reference population or control population. Such a reference population can be used for comparison of an unknown sample. One skilled in the art can determine an
15 appropriate reference population based on the particular application of the methods of the invention. The methods of the invention can be used to generate a differential polypeptide profile that identifies the differences in polypeptide expression between two samples, for example, a
20 normal and disease state. The size of the reference population depends on the criteria used to select reference individuals. Depending on the selection criteria and particular application of the methods of the invention, a reference population can be a relatively small number to a
25 large number of individuals, including thousands of individuals.

The large-scale analysis of samples from patients having specifically diagnosed diseases or exhibiting signs or symptoms of a disease is useful for identifying clinical
30 markers or constellations of markers for the respective

conditions. Samples from an individual having a disease can be used to generate a qualitative and/or quantitative polypeptide identification index for that disease. Similarly, the comparative analysis of polypeptides

5 contained in samples from patients undergoing therapeutic treatment can be used to identify diagnostic markers or constellation of markers indicating the success or failure of the treatment. The methods are also applicable to the analysis of such samples on a systematic, population-wide

10 scale for the discovery or screening of markers or constellations of markers useful for indicating the predisposition of individuals for certain clinical conditions.

The invention further provides a method for

15 generating a polypeptide identification index. The method includes steps of (a) determining a set of two or more characteristics associated with a first polypeptide, or a peptide fragment thereof, one of the characteristics being the mass of a peptide fragment of the polypeptide, the

20 peptide fragment mass being determined by mass spectrometry; (b) repeating step (a) for a second polypeptide; (c) optionally determining one or more additional characteristics associated with the first and second polypeptides, wherein the determined characteristics are

25 sufficient to distinguish the first and second polypeptides, thereby generating a polypeptide identification index for the first and second polypeptides. The method can further comprise repeating steps (a) through (c) one or more times for a different polypeptide, wherein the determined

30 characteristics are sufficient to distinguish each of the polypeptides, thereby generating a polypeptide

identification index for each of the polypeptides. As with determining characteristics of a polypeptide, the polypeptide identification can be determined with any of the methods disclosed herein for determining characteristics associated with a polypeptide.

The methods of the invention for generating a polypeptide identification index can involve the determination of polypeptide or fragment mass at an accuracy in ppm of greater than 1 part per million or even lower accuracy (higher ppm), if desired. Although a polypeptide identification index can contain a polypeptide amino acid sequence, it is not required that a polypeptide or fragment be sequenced for practicing the methods of the invention for generating a polypeptide identification index.

The methods of the invention involve determining characteristics associated with a polypeptide. A sample containing a polypeptide can be as simple as an isolated polypeptide mixture containing a polypeptide or as complex as sample containing all of the polypeptides expressed in an organism. Furthermore, a sample can be fractionated, if desired, using the methods disclosed herein.

A polypeptide can be in a sample isolated from a variety of sources. For example, a polypeptide sample can be prepared from any biological fluid, cell, tissue, organ or portion thereof, or any species of organism. A sample can be present in an individual and obtained or derived from the individual. For example, a sample can be a histologic section of a specimen obtained by biopsy, or cells that are placed in or adapted to tissue culture. A sample further

can be a subcellular fraction or extract. A sample can be prepared by methods known in the art suitable for maintaining polypeptide solubility, such as those described herein.

5 A specimen refers specifically to a sample obtained from an individual. A specimen can be obtained from an individual as a fluid or tissue specimen. For example, a tissue specimen can be obtained as a biopsy such as a skin biopsy, tissue biopsy or tumor biopsy. A fluid
10 specimen can be blood, serum, urine, saliva, cerebrospinal fluid or other bodily fluids. A fluid specimen is particularly useful in methods of the invention since fluid specimens are readily obtained from an individual. Methods for collection of specimens are well known to those skilled
15 in the art (see, for example, Young and Bermes, in Tietz Textbook of Clinical Chemistry, 3rd ed., Burtis and Ashwood, eds., W.B. Saunders, Philadelphia, Chapter 2, pp. 42-72 (1999)).

 A polypeptide to be used in the methods of the
20 invention can be obtained from a source such as a cell, tissue, organ or organism. A variety of methods are known in the art for lysing a cell. Cells can be lysed, for example, by denaturants, one or more cycles of freezing and thawing, and sonication. Following lysis, the polypeptide
25 mixture can be subjected to a fractionation to remove, for example, nucleic acid or lipid, or to remove intact subcellular fractions or organelles. Methods of lysing and fractionating cells are well known to those skilled in the art (see Scopes, *supra*, 1993).

For identification of a polypeptide, a sample or specimen can be contained in a buffer suitable for maintaining polypeptide solubility such as, for example, a buffer containing a detergent, including denaturants such as sodium dodecyl sulfate (SDS). Denaturants useful for solubilizing polypeptides include, for example, guanidine-HCl, guanidine-isothiocyanate and urea. In the case of guanidine-isothiocyanate, as with treatment with any reagent that can covalently modify a polypeptide, such reagents can be used so long as the polypeptide identification index to which the sample is to be compared has been prepared in substantially the same manner as the sample sufficient for comparison of the same polypeptide. Other denaturants well known in the art can be similarly used for solubilizing polypeptides. In addition, reducing agents such as dithiothreitol (DTT), dithioerythritol (DTE), or mercaptoethanol can be included.

The methods of the invention can optionally involve protein fractionation steps. Protein fractionation refers to any method useful for removing one or more polypeptides from a polypeptide population. Fractionation can include, for example, a centrifugation step that separates soluble from insoluble components, a method of electrophoresis, and a method of chromatography, or any of the methods disclosed. For chromatographic separation, a wide variety of chromatographic media well known in the art can be used to separate polypeptide populations. For example, polypeptides can be separated based on size, charge, hydrophobicity, binding to particular dyes and other moieties associated with chromatographic media. Size exclusion, gel filtration and gel permeation resins are

useful for polypeptide separation based on size. Examples of chromatographic media for charge-based separation are strong and weak anion exchange and strong and weak cation exchange resins. Hydrophobic or reverse phase
5 chromatography can also be used.

Affinity chromatography can also be used including, for example, dye-binding resins such as Cibacron blue, substrate analogs, including analogs of cofactors such as ATP, NAD, and the like, ligands, specific antibodies,
10 either polyclonal or monoclonal, and the like. An exemplary affinity resin includes affinity resins that bind to specific moieties that can be incorporated into a polypeptide such as an avidin resin that binds to a biotin tag on a polypeptide, as disclosed herein. The resolution
15 and capacity of particular chromatographic media are known in the art and can be determined by those skilled in the art. The usefulness of a particular chromatographic separation for a particular application can similarly be assessed by those skilled in the art.

Those of skill in the art will be able to
20 determine the appropriate chromatography conditions for a particular sample size or composition and will know how to obtain reproducible results for chromatographic separations under defined buffer, column dimension, and flow rate
25 conditions. All protein fractionation methods can optionally include the use of an internal standard for assessing the reproducibility of a particular chromatographic application. Appropriate internal standards will vary depending on the chromatographic medium. Those

skilled in the art will be able to determine an internal standard applicable to a method of chromatography.

Polypeptide tagging is useful in the methods of the invention for reducing polypeptide sample complexity, providing a database search constraint, and enabling quantitative polypeptide comparisons. The complexity of a polypeptide sample can be reduced by tagging a polypeptide with an affinity tag that can be used for isolating a subpopulation of polypeptides that contain the tag. For example, a population of polypeptides and fragments can be labeled on a relatively rare amino acid, such as cysteine, and a subpopulation of polypeptides and fragments containing the tag can be isolated. The subpopulation of polypeptides and fragments isolated in this manner will thus contain a known amino acid. As described herein, a known amino acid constitutes a partial amino acid composition which is useful for constraining a database search. Quantitative polypeptide comparisons can be performed by differentially tagging two polypeptides or polypeptide populations. The ICAT™ type affinity reagent, described in more detail below, is particularly useful for this purpose, although any other method of polypeptide tagging can be similarly applied to polypeptide comparisons.

Polypeptide tagging can be performed using a variety of methods known in the art. A reagent for polypeptide tagging or modification can contain various components that are separated by linker regions. Components of a polypeptide tagging reagent can include a reactive group that modifies a specific chemical group of a polypeptide, a moiety that can be detected, such as by mass

spectrometry, and an affinity tag to be used for polypeptide isolation. Two examples of polypeptide tagging reagents, ICAT™ type and IDEnT, are described in detail below, although any type of polypeptide tag can be used, if
5 desired.

The methods of the invention for quantitatively comparing two polypeptide populations involve the use of the isotope-coded affinity tag (ICAT™) type method (Gygi et al.,
10 Nature Biotechnol. 17:994-999 (1999) which is incorporated herein by reference). An ICAT™ type reagent can additionally be useful for polypeptide tagging applications that do not involve quantitative comparisons. The ICAT™ type reagent method uses an affinity tag that can be
15 differentially labeled with an isotope that is readily distinguished using mass spectrometry, for example, hydrogen and deuterium. The ICAT™ type affinity reagent consists of three elements, an affinity tag, a linker and a reactive group.

20

One element of the ICAT™ type affinity reagent is an affinity tag that allows isolation of peptides coupled to the affinity reagent by binding to a cognate binding partner of the affinity tag. A particularly useful affinity tag is
25 biotin, which binds with high affinity to its cognate binding partner avidin, or related molecules such as streptavidin, and is therefore stable to further biochemical manipulations. Any affinity tag can be used so long as it provides sufficient binding affinity to its cognate binding
30 partner to allow isolation of peptides coupled to the ICAT™ type affinity reagent. An affinity tag can also be used to isolate a tagged peptide with magnetic beads or other

magnetic formats suitable to isolate a magnetic affinity tag. In the ICAT™ type reagent method, or any other method of affinity tagging a peptide, the use of covalent trapping can be used to bind the tagged peptides to a solid support,
5 if desired.

A second element of the ICAT™ type affinity reagent is a linker that can incorporate a stable isotope. The linker has a sufficient length to allow the reactive group to bind to a specimen polypeptide and the affinity tag
10 to bind to its cognate binding partner. The linker also has an appropriate composition to allow incorporation of a stable isotope at one or more atoms. A particularly useful stable isotope pair is hydrogen and deuterium, which can be readily distinguished using mass spectrometry as light and
15 heavy forms, respectively. Any of a number of isotopic atoms can be incorporated into the linker so long as the heavy and light forms can be distinguished using mass spectrometry. Exemplary linkers include the 4,7,10-trioxa-1,13-tridecanediamine based linker and its related
20 deuterated form, 2,2',3,3',11,11',12,12'-octadeutero-4,7,10-trioxa-1,13-tridecanediamine, described by Gygi et al. (*supra*, 1999). One skilled in the art can readily determine any of a number of appropriate linkers useful in an ICAT™ type affinity reagent that satisfy the above-described
25 criteria.

The third element of the ICAT™ type affinity reagent is a reactive group, which can be covalently coupled to a polypeptide in a specimen. Methods for modifying side chain amino acids in polypeptides are well known to those
30 skilled in the art (see, for example, Glazer et al.,

Laboratory Techniques in Biochemistry and Molecular Biology:
Chemical Modification of Proteins, Chapter 3, pp. 68-120,
Elsevier Biomedical Press, New York (1975); Pierce Catalog
(1994), Pierce, Rockford IL). Any of a variety of reactive
5 groups can be incorporated into an ICAT™ type affinity
reagent so long as the reactive group can be covalently
coupled to a polypeptide. For example, a polypeptide can be
coupled to the ICAT™ type affinity reagent via a sulfhydryl
reactive group, which can react with free sulfhydryls of
10 cysteine or reduced cystines in a polypeptide. An exemplary
sulfhydryl reactive group includes an iodoacetamido group,
as described in Gygi et al. (*supra*, 1999). Other exemplary
sulfhydryl reactive groups include maleimides, alkyl and
aryl halides, α -haloacyls and pyridyl disulfides. If
15 desired, the polypeptides can be reduced prior to reacting
with an ICAT™ type affinity reagent, which is particularly
useful when the ICAT™ type affinity reagent contains a
sulfhydryl reactive group.

A reactive group can also react with amines such
20 as Lys, for example, imidoesters and N-hydroxysuccinimidyl
esters. A reactive group can also react with carboxyl
groups found in Asp or Glu, or the reactive group can react
with other amino acids such as His, Tyr, Arg, and Met. A
reactive group can also react with a phosphate group for
25 selective labeling of phosphopeptides, or with other
covalently modified peptides, including glyopeptides,
lipopeptides, or any of the covalent polypeptide
modifications disclosed herein. One skilled in the art can
readily determine conditions for modifying specimen
30 molecules by using various reagents, incubation conditions
and time of incubation to obtain conditions optimal for

modification of specimen molecule for use in methods of the invention.

The ICAT™ type reagent method is based on derivatizing a specimen molecule such as a polypeptide with
5 an ICAT™ type affinity reagent. A control reference specimen and a specimen from an individual to be tested are differentially labeled with the light and heavy forms of the ICAT™ type affinity reagent. The derivatized specimens are combined, and the derivatized molecules cleaved to generate
10 fragments. For example, a polypeptide molecule can be enzymatically cleaved with one or more proteases into peptide fragments. Exemplary proteases useful for cleaving polypeptides include trypsin, chymotrypsin, pepsin, papain, *Staphylococcus aureus* (V8) protease, and the like.
15 Polypeptides can also be cleaved chemically, for example, using CNBr, acid or other chemical reagents.

Once cleaved into fragments, the tagged fragments derivatized with the ICAT™ type affinity reagent are isolated via the affinity tag, for example, biotinylated
20 fragments can be isolated by binding to avidin in a solid phase or chromatographic format. If desired, the isolated, tagged fragments can be further fractionated using one or more alternative separation techniques, including ion exchange, reverse phase, size exclusion affinity
25 chromatography and the like, or electrophoretic methods, including isoelectric focusing. For example, the isolated, tagged fragments can be fractionated by high performance liquid chromatography (HPLC), including microcapillary HPLC.

The fragments are analyzed using mass spectrometry (MS). Because the specimen molecules are differentially labeled with light and heavy affinity tags, the peptide fragments can be distinguished on MS, allowing a side-by-side comparison of the relative amounts of each peptide fragment from the control reference and test specimens. If desired, MS can also be used to sequence the corresponding labeled peptides, allowing identification of molecules corresponding to the tagged peptide fragments.

10 An advantage of the ICAT™ type reagent method is that the pair of peptides tagged with light and heavy ICAT™ type reagents are chemically identical and therefore serve as mutual internal standards for accurate quantification (Gygi et al., *supra*, 1999). Using MS, the ratios between
15 the intensities of the lower and upper mass components of pairs of heavy- and light-tagged fragments provides an accurate measure of the relative abundance of the peptide fragments. Thus, the ICAT™ type reagent method can be conveniently used to identify differentially expressed
20 polypeptides, if desired.

An IDEnT reagent can be used to modify a polypeptide by introducing an isotopic tag at a specific protein functional group. An exemplary IDEnT reagent is described in Goodlett et al., *supra*, 2000. An IDEnT reagent
25 contains at least one element with an isotopic distribution that creates a unique signature in a mass spectrometer. For example, an IDEnT reagent can contain chlorine, deuterium, or another element, including a radioactive element. An IDEnT reagent can be designed to bind to a low abundance
30 amino acid in a polypeptide, such as cysteine. The labeling

of a polypeptide with an IDEnT tag can be applied to the methods of the invention by providing a constraint for searching a polypeptide identification index with polypeptide fragment masses.

5 Protein cleavage or fragmentation is useful in the methods of the invention for providing a constraint for database searching. Polypeptide fragmentation can be sequence-specific or non-specific. Sequence-specific polypeptide cleavage provides the advantage of obtaining
10 polypeptide fragments that contain known amino acids which can be used to constrain a database search. Examples of reagents useful for performing non-specific polypeptide cleavage are papain, pepsin and protease Sg. These proteases can be used to achieve a desired degree of protein
15 fragmentation, such as, for example, the generation of about two to four polypeptide fragments from a polypeptide by altering the reaction conditions. Conditions for using these proteases are well known in the art. Examples of reagents useful for performing sequence-specific polypeptide
20 cleavage are trypsin, V-8 protease, o-iodosobenzoic acid, cyanogen bromide and acid.

The invention also provides a polypeptide identification index for identifying a polypeptide from a population of polypeptides. The index comprises an
25 annotated set of characteristics associated with polypeptides in the index, one of the characteristics being the mass of a fragment of the polypeptide. The characteristics are sufficient to distinguish one of the polypeptides from other polypeptides in the index. The
30 polypeptide index can comprise characteristics for 2 or

more, 3 or more, 5 or more, 10 or more, 20 or more, 50 or more, 100 or more, 200 or more, 500 or more, 1000 or more, 2000 or more, 5000 or more, or even 10,000 or more polypeptides. A polypeptide identification index can also
5 include substantially all of the polypeptides in a sample. For example, a polypeptide identification index can include substantially all of the polypeptides expressed in a genome, such as a viral, bacterial, plant or animal genome, including a mammalian genome such as human, non-human
10 primates, mouse, rat, bovine, goat, rabbit or other mammalian species. The number of polypeptides in a polypeptide identification index will depend on the needs of the user and will vary depending on the source of the sample to be used to identify polypeptides and the complexity of
15 polypeptide expression in the sample.

The polypeptide identification index can be directed to a whole organism or to particular tissues or cells in an organism or to specific subcellular fractions, for example, organelles, as desired. Accordingly, similar
20 to the reduction in complexity applied to a sample to be tested, a polypeptide identification index directed to a particular target such as an organism, tissue, cell or subcellular fraction, can be useful for simplifying a search for identification of a particular polypeptide in a
25 particular application. For example, in a particular diagnostic application where expression of a particular polypeptide or group of polypeptides, or the amount of expression of the polypeptides, is correlated with a particular condition such as a disease condition, the use of
30 a polypeptide identification index directed to a relevant target can be used. For example, if a group of nuclear

proteins are known to be overexpressed in a cancer cell, the use of a polypeptide identification index directed to nuclear proteins can be used to test for overexpression of the nuclear proteins in a sample from an individual using
5 the quantitative methods disclosed herein. Moreover, the generation of a targeted polypeptide identification index and comparison to a relevant disease sample can be used to identify aberrantly expressed polypeptides, which in turn can be used in diagnostic applications, as disclosed herein.

10 The invention additionally comprises a polypeptide identification index comprising an annotated set of characteristics associated with polypeptides of the index comprising two or more characteristics associated with polypeptides of the index, or a fragment thereof, one of the
15 characteristics being the mass of a fragment of the polypeptide, and wherein the mass is determined at an accuracy in ppm of greater than 1 ppm.

If desired, a polypeptide identification index can be conveniently stored on a computer readable medium.
20 Accordingly, the invention provides a computer readable medium comprising an invention polypeptide identification index, for example, an annotated polypeptide index. Such a computer readable medium comprising a polypeptide identification index is useful for comparing the
25 characteristics of a polypeptide with the polypeptide identification index, which can be conveniently performed on a computer apparatus. The use of a computer apparatus is convenient since a polypeptide identification index can be conveniently stored and accessed for comparison to
30 characteristics and/or quantitative amounts of a polypeptide

in a sample. A polypeptide identification index can be conveniently accessed using appropriate hardware, software, and/or networking, for example, using hardware interfaced with networks, including the internet.

5 By using various hardware, software and network combinations, the methods of the invention including the step of comparing the characteristics determined for a polypeptide to a polypeptide identification index can be conveniently performed in a variety of configurations.

10 Accordingly, the invention additionally provides a computer apparatus for carrying out computer executable steps corresponding to steps of invention methods. For example, a single computer apparatus can contain instructions for carrying out the computer executable step(s) of comparing

15 characteristics determined for polypeptide to a polypeptide identification index, a polypeptide identification index, and instructions for determining whether the characteristics determined for the polypeptide correspond to one or more polypeptides in the polypeptide identification index.

20 Alternatively, the computer apparatus can contain instructions for carrying out the steps of an invention method while the polypeptide identification index is stored on a separate medium. In addition, instructions for determining whether a polypeptide corresponds to one or more

25 polypeptides in the polypeptide identification index can be contained on a separate computer apparatus or separate medium, or combined with the computer apparatus containing the computer executable steps of the method and/or the database on a separate medium. Such a separate computer

30 readable medium can be another computer apparatus, a storage

medium such as a floppy disk, Zip disk or a server such as a file-server, which can be accessed by a carrier wave such as an electromagnetic carrier wave. Thus, a computer apparatus containing a polypeptide identification index or a file-
5 server on which the polypeptide identification index is stored can be remotely accessed via a network such as the internet. One skilled in the art will know or can readily determine appropriate hardware, software or network interfaces that allow interconnection of an invention
10 computer apparatus.

It is understood that modifications which do not substantially affect the activity of the various embodiments of this invention are also included within the definition of the invention provided herein. Accordingly, the following
15 examples are intended to illustrate but not limit the present invention.

EXAMPLE I

Generation of an Annotated Polypeptide Index

This example describes the generation of an
20 annotated polypeptide index and use of the annotated polypeptide index to identify a polypeptide in a sample.

The elements of an annotated polypeptide (AP) index, also referred to as an annotated peptide tag (APT) index or database, are the sequences of essentially all the
25 peptides or selected peptides with specific structural features that are generated by sequence specific chemical or enzymatic fragmentation of the proteins produced by the species, cell or tissue under investigation. Each peptide

is annotated with attributes, or characteristics, that are easily determined experimentally and that permit the unambiguous correlation between the annotated peptide and the protein from which the peptide originated.

5 The generation of an exemplary AP index can involve the following specific steps: harvest proteins; label proteins with an isotope coded affinity tag (ICAT™) type reagent; fractionate proteins by molecular weight; digest proteins with a protease, for example, trypsin, to
10 generate peptides; separate peptides by chromatography, for example, ion exchange chromatography; purify each ion exchange fraction by affinity chromatography, for example, based on the ICAT™ type affinity tag; analyze each affinity chromatography fraction by LC/MS/MS or CE/MS/MS; identify
15 essentially all expressed proteins via a database search of individual MS/MS peptide spectra; and generate a database of annotated peptide tags that constitute a unique bar code for an individual peptide based on measured physicochemical properties and thus the parent protein of that peptide.

20 The AP index can be generated as follows: (i) protein sample preparation; (ii) protein tagging; (iii) optional protein fractionation; (iv) protein fragmentation; (v) peptide separation; (vi) affinity isolation of tagged peptides; (vii) high resolution peptide separation; (viii)
25 database searching; (ix) AP index (APT database) construction.

(i) Protein Sample Preparation. Protein samples for which quantitative proteome analysis is to be performed, for example, cells, tissues, subcellular fractions, body

fluids, cellular secretions, and the like, are isolated from the respective sources using standard protocols for maintaining the solubility of all the proteins.

(ii) Protein tagging. The proteins in the sample
5 are completely denatured, reduced, and the all the
sulfhydryl groups representing the side chains of reduced
cysteine residues are covalently derivatized with the light
or heavy form, respectively, of sulfhydryl-specific ICAT™
type reagents using the conditions described previously
10 (Gygi et al., Nature Biotechnol. 17:994-999 (1999)) (see
Figure 2. While cysteine tagging is a particularly useful
implementation of the method, any other chemical reaction
with specificity for a chemical group in the protein can
also be applied.

(iii) Optional Protein Fractionation. The mixture
15 of tagged proteins is fractionated using any one of the
known standard protein separation procedures. The applied
procedure is reproducible, maintains the proteins in
solution, and has a high sample capacity. A particularly
20 useful method is preparative sodium dodecyl sulfate-
polyacrylamide gel electrophoresis (SDS-PAGE).

(iv) Protein Fragmentation. The proteins in the
sample mixture, or the proteins contained in each fraction
if optional sample fractionation is employed, are subjected
25 to sequence specific cleavage. A preferred method is
tryptic cleavage.

(v) Peptide Separation. The resulting peptide
mixtures are subjected to a first dimension peptide

separation. The peptide separation method has a high sample capacity, at least moderate resolving power, and generates highly reproducible separation patterns, irrespective of the complexity of the sample applied. A particularly useful
5 first dimension separation method is cation ion exchange chromatography.

(vi) Affinity Isolation of Tagged Peptides.

Peptides tagged with the ICAT™ type reagent, i.e., cysteine containing peptides, are isolated from each chromatographic
10 fraction using avidin or streptavidin affinity chromatography. A particularly useful affinity medium is monomeric avidin immobilized on polymer beads. If ICAT™ type reagents with affinity tags different from biotin are used, affinity media complementary to that tag are used.

15 (vii) High Resolution Peptide Separation. A particularly useful method for high resolution peptide separation is liquid chromatography ESI-MS/MS. The peptide mixtures eluted from the affinity chromatography columns are individually analyzed by automated LC-MS/MS using capillary
20 reversed phase chromatography as the separation method (Yates et al., Methods Mol. Biol. 112:553-569 (1999)) and data dependent CID with dynamic exclusion (Goodlett et al., Anal. Chem. 15:1112-1118 (2000)) as the mass spectrometric method.

25 (viii) Database Searching. The sequence of all the peptides for which suitable CID spectra are obtained is determined by searching a sequence database from the species under investigation. A particularly useful sequence database is a database containing all the complete protein

sequences that can be potentially expressed by the species under examination. The sequence database search program is the SEQUEST® program (Eng, J. et al., J. Am. Soc. Mass. Spectrom. 5:976-989, (1994)) or a program with similar
5 capabilities.

(ix) AP index (APT database) Construction. The sequences of all the peptides that have been identified by the procedure described above are entered in a database and annotated with the characteristics, or attributes, that were
10 generated during steps (i)-(viii) above. These characteristics, or attributes, include, but are not limited to: partial amino acid composition (such as the presence of a cysteine residue in each selected peptide; see Goodlett et al., *supra*, 2000); the approximate molecular mass of the
15 parent protein (as determined by the optional SDS-PAGE fractionation); the order of elution or elution time from the cation ion exchange column; and the elution time from the reversed-phase column. Collectively, these attributes are unique for every peptide in the database akin to a bar
20 code for each peptide. Therefore, if the same bar code is being determined from the peptides generated in subsequent experiments, they will uniquely identify the peptides generated by the experiment, simply by correlating the bar codes generated by the experiment with the bar codes present
25 in the AP index (APT database).

For correlation of polypeptides with the AP index (APT database), the peptide samples generated for quantitative proteome analysis by the method described above are generated, treated and processed precisely like the
30 peptides generated for the AP index (APT database), with the

following exceptions. (i) The proteins in the two (or more) samples to be compared are labeled with differentially isotopic labeled ICAT™ type reagents. Two or several chemically identical but differentially isotopically labeled ICAT™ type reagents can be used at this step. (ii) The generated peptides are analyzed by single stage mass spectrometry only, rather than by MS/MS. Mass analyzers will generally have a high mass accuracy, high sensitivity and high mass resolution. Instruments with these characteristics include, but are not limited to, MALDI-TOF mass spectrometers, ESI-TOF mass spectrometers and Fourier transform ion cyclotron mass analyzers (FT-ICR-MS). The attributes determined from each peptide by this process (all the attributes described above or a selection thereof) are translated into a bar code for each peptide, and the experimentally determined bar code is correlated with the bar codes from the AP index (APT database), resulting in the unambiguous identification of the peptide and therefore the protein from which the peptide originated. Subsequently, the mass spectra are examined for pairs of peptide ions that co-fractionated throughout the process and that have a mass difference that precisely corresponds to the mass difference encoded in the ICAT™ type reagent used. The relative signal intensities of the two peaks indicate the relative abundance of the peptides and therefore the relative abundance of the corresponding proteins initially present in the sample. Consequently, the correlation of the experimentally determined data with the AP index (APT database) allows quantification and identification of the proteins in the samples analyzed.

EXAMPLE II**Generation of a Yeast Annotated Polypeptide Index**

This example describes the generation of an annotated polypeptide index for yeast.

5 At least 5 mg of total protein was estimated to be required at current mass spectrometer sensitivity to detect low abundance proteins using the LC/LC/MS/MS method (Gygi et al., Proc. Natl. Acad. Sci. USA 97:9390-9395 (2000)), and this amount was essentially experimentally confirmed. Gygi
10 et al., *supra*, 2000 also demonstrated that the "binning" process is adequate for the detection of low abundance proteins and has sufficient sample capacity to accommodate the relatively large amounts of total sample.

 For the construction of the database, the
15 following procedure is used. For protein labeling, a protein sample is generated in 0.5% SDS, 50 mM Tris, pH 8.3, 5 mM ethylenediaminetetraacetic acid (EDTA) at a protein concentration of 5 mg/ml. A total of 25 mg of total yeast protein is used. Once proteins are in solution, the SDS
20 concentration is lowered by diluting the sample 1:10 with water and adding EDTA to maintain a 5 mM EDTA concentration. The final concentration is 0.05% SDS, 5 mM Tris, 5 mM EDTA. The sample is then boiled for 3-5 min at 100°C and then chilled. Reduction of disulfide bonds is accomplished by
25 adding sufficient Tributylphosphine (TBP) to achieve 5 mM in the sample solution. This is followed by an incubation of the sample at 37°C for 30 min. To the reduced sample, the alkylating reagent (e.g. ICAT™ type reagent) is added at an estimated 5x molar excess over the SH groups present in the

sample. The alkylation reaction is allowed to proceed in darkness for 90 min.

For protein separation, the reduced and alkylated sample is added with .2 volume of 5x SDS gel sample buffer and boiled for 5 min. The cooled sample is then applied to a preparative SDS gel with the dimensions 20cm x 20cm x 1.5 mm. After electrophoresis, the gel is sliced perpendicular to the electrophoresis dimension into 10 strips of equal size. These strips represent 10 size bins for the intact proteins. The proteins in the gel strips are then subjected to in-gel digestion using standard protocols.

For peptide separation, the peptides that are extracted from the gel slices are subjected to three sequential chromatographic separations. First, they are separated by cation exchange chromatography. Second, the biotinylated peptides are isolated by avidin chromatography. Third, peptides are further fractionated by capillary reverse-phase chromatography.

For cation exchange chromatography, a cation exchange HPLC column is used (PolyLC Inc., 2.1 mm x 20 cm, 5 μ m particles, 300 Å pore size, Polysulfoethyl A strong cation exchange material). The following buffers are used: Buffer A, 10 mM KH_2PO_4 , 25% CH_3CN , pH 3.0; Buffer B, 10 mM KH_2PO_4 , 25% CH_3CN , 350 mM KCl, pH 3.0. The following gradient is run:

| Time (min) | %B |
|------------|-----|
| 0 | 0 |
| 30 | 25 |
| 50 | 100 |

The flow rate is 200 μ L per minute. Fractions are collected at 1-2 minute intervals. Anywhere from about 200 microgram up to about 5 mg of digested, ICATTM labeled total protein is loaded on this column, usually using a 2 mL sample loop. It is important to acidify the samples down to pH 3.0 or below before loading onto the cation exchange column because peptides will not be fully charged at higher pH values and can possibly not stick to the column. The gradient shown is designed to spread out the elution of doubly-charged peptides as much as possible, with these peptides usually eluting starting at about 8-9 minutes into the run until approximately minutes 15-16, after which triply charged peptides begin to elute. 30 fractions are collected over the duration of the gradient.

For avidin affinity chromatography, an Ultralink Monomeric Avidin (Pierce, cat # 53146) is used. A small piece of glass wool is packed into the neck of a glass pipette tube. 400 μ l of avidin chromatography material is packed into the tube (slurry comes at 50% dilution, so 800 μ l of 50% slurry is added in order to get 400 μ l avidin chromatography material). The beads are allowed to settle, and the column is washed with 2 X PBS to bring the beads down off the side of the tube. The packed column is washed through with 30% Acetonitrile (ACN) with 0.4% Trifluoroacetic acid (TFA) until the flow-through pH changes to ~1, and then another column volume of the ACN/TFA is washed through. This acidic step is to get rid of polymers associated with the beads. The column is washed with 2 X PBS pH 7.2 until the pH is ~7.2. Thereafter, the column is washed through with 3 more column volumes (1200 μ l) of same buffer.

The column is washed with 3-4 column volumes of biotin blocking buffer (2 mM d-biotin in PBS). This biotin blocks the more retentive avidin sites on the column, ensuring recovery of the sample from the remaining binding sites later on.

Loosely bound biotin is washed off with ~6 column volumes (2,400 μ l) of regeneration buffer (100 mM glycine, pH 2.8), until the flow-through pH changes to ~2.8. This glycine solution is sterilized by autoclaving before use and stored at 4°C, where it will last one week.

The column is washed with 6 column volumes of 2 X PBS to return column to proper pH (~7.2). The flow-through pH is monitored. The peptide samples consisting of individual or pooled ion exchange column fractions are then applied to the column and incubated in column for ~20 min.. Unbound material is then washed from the column by applying 5 column volumes of 2 X PBS, pH 7.2, and fractions are collected. The column is further washed through with 5 X column volumes of 1 X PBS (this step is to reduce the salt concentration), and 6 column volumes of 50 mM AMBIC, pH 8.3, with 20 % methanol (MeOH) while continuing to collect fractions (50 mM AMBIC is to bring the salt concentration down; MeOH is to get rid of hydrophobic peptides). Biotinylated peptides are eluted with elution buffer (30% ACN / 0.4% TFA) and collected manually in a glass tube. These samples are further separated by capillary reverse-phase chromatography.

For reverse-phase chromatography, purified biotinylated peptides are separated by reverse-phase

capillary chromatography using standard protocols. The solvent gradient is chosen so that the peptides elute over 60 min. If 1 min fractions are collected and analyzed in the mass spectrometer, 60 bins are created by this
5 procedure.

For mass analysis and sequencing, biotinylated peptides eluting from the RP-columns are analyzed by ESI-MS/MS for the generation of the database and by ESI-MS for database searching. For the database construction, an ion
10 trap mass spectrometer is used with a mass accuracy of approximately 1 mass unit. For the mass measurement for database searching, an ESI-TOF mass spectrometer is used with a mass accuracy exceeding 50 ppm and a resolution exceeding 10.000. That means that a peptide at 1000 mass
15 units can be distinguished from a peptide at 1000.1 mass units. If a mass range for tryptic peptides is assumed to be between 500 and 3000 mass units, a mass spectrometer at that performance would generate 25.000 bins.

In the case of proteome analysis of the yeast *S. cerevisiae*, the genome of this organism contains
20 approximately 6200 ORF's. The yeast proteome therefore is expected to be approximately 6200 different proteins, disregarding differentially modified forms of the same protein. Tryptic digestion of a yeast proteome would yield
25 approximately 350,000 peptides if empirically derived specificity rules for trypsin are applied. This sample complexity is reduced to approximately 35,000 peptides if only the cysteine-containing peptides are extracted, based on the chemical derivatization with the ICAT™ type reagents.
30 The total number of bins available from the procedure

described above is $10 \times 30 \times 60 \times 25.000 = 4,5 \times 10^8$ and therefore by far exceeds the number of peptides expected from a total yeast proteome analysis. It is therefore expected that for a sample of the complexity of a yeast extract, the procedure
5 for the generation of database search data can be simplified. As an example, the gel electrophoresis sizing step for proteins can be optionally eliminated.

Neither the procedure for the generation of the data to be entered into the database nor the procedure for
10 the generation of data to search the database are fixed. Therefore, for optimization, depending on the degree of sample complexity, the number of bins available can be easily adjusted. Generally, the number of bins chosen for the generation of the database is high, whereas the number
15 of bins for generating the database search data would be chosen as low as possible to maximize the sample throughput. The number of bins available can be easily adapted in various ways. Firstly, the inclusion of additional orthogonal separation dimensions can be considered for
20 proteins or peptides. For protein separation, isoelectric focusing, ion exchange chromatography, hydroxylapatite chromatography, or similar electrophoretic or chromatographic techniques can be included. For peptide
25 separation, separation based on peptide size or capillary electrophoresis methods can be included.

Secondly, the separation range for the separation methods described above can be extended. Protein sizing can be extended by using gradient gels or longer gels with extended separation range. For the chromatographic peptide
30 separation methods, the number of bins can be easily

expanded by generating extended, shallower gradients and/or by sampling more frequently. Finally the number of bins is critically dependent on the resolution and mass accuracy of the mass analyzer used. Adding a mass analyzer with higher
5 performance will decrease the number of bins provided by the separation methods employed in the procedure.

Throughout this application various publications have been referenced within parentheses. The disclosures of these publications in their entireties are hereby
10 incorporated by reference in this application in order to more fully describe the state of the art to which this invention pertains.

Although the invention has been described with reference to the disclosed embodiments, those skilled in the
15 art will readily appreciate that the specific experiments detailed are only illustrative of the invention. It should be understood that various modifications can be made without departing from the spirit of the invention. Accordingly, the invention is limited only by the following claims.